Texto! octobre 2008, vol. XIII, n°4

Corpus, quantification et typologie textuelle

Sylvain Loiseau
LIMSI-CNRS

« la typologie ne saurait se contenter d'étudier
la grammaire. Le discours est une réalité
fondamentale, et c'est au croisement des deux
que devraient se situer les études typologiques. »
(Hagège, 2005 : 68)

Résumé

Cet article examine les conséquences pour la description de normes linguistiques des nouveaux observables que permettent de construire les linguistiques de corpus. Nous formulons l'hypothèse que des corpus complexes articulant plusieurs niveaux de descriptions permettent de caractériser les textes, au-delà de profils morpho-syntaxiques ou lexicaux, par des observables fondés sur des catégories de la linguistique textuelle. L'enjeu est de pouvoir articuler méthodes quantitatives et linguistique textuelle pour renouveler les capacités descriptives. Dans le cadre d'une linguistique des normes, cette articulation peut contribuer à une typologie textuelle, qui doit être rapprochée des typologies linguistiques.

Abstract

In this paper, we examine the construction of observables, in quantitative corpus linguistics, for the description of linguistic genres and discourses and for textual typology. We propose a hypothesis according to which texts may be characterized through correlations between levels of description, rather that using one isolated level of description at a time. The systemic relations between levels of description allows us to use categories of textual linguistics in quantitative analysis. New contextualities have to be considered: contextuality as interaction between levels of description, and contextuality

as interaction between genres, discourses, and idiolects in each text. Coupling textual linguistics and quantitative corpus linguistics raises the issue of the relation between textual typology and linguistic typology.

Mots clefs

Typologie textuelle, genre, linguistique textuelle, linguistique de corpus, norme, linguistique quantitative.

Key words

Textual typology, genre, textual linguistics, corpus linguistics, norm, quantitative linguistics.

Introduction

De nombreux travaux en description de « types de texte », dans le cadre des linguistiques de corpus, ont rendu possible une nouvelle forme de description de la variation interne d'une langue et une nouvelle forme de typologie, fondée sur des données quantitatives[1]. Les enjeux de l'identification et de la description de types de textes sont devenus essentiels pour les « traitements automatiques » des langues, puisqu'ils permettent de concevoir des outils non plus pour une langue, mais pour un type de texte, et d'accroître ainsi leur qualité. Mais ces enjeux sont tout autant essentiels pour la description linguistique elle-même, puisqu'ils font apparaître de nouveaux observables et permettent de reconsidérer des questions comme celles de la relation entre système et variation, de la distinction des niveaux de description, ou de la relation entre description « quantitative » et « qualitative ».

Or, la plupart des travaux en typologie textuelle s'appuient sur un seul niveau de description – généralement le niveau

1 Je remercie J.-B. Berthelin et C. Grouin qui ont discuté et critiqué les versions préliminaires de cet article. Ce travail a bénéficié d'un financement de l'Agence National de la Recherche (ANR), pour le projet C-Mantic (ANR-07-MDCO-002).

morphosyntaxique – en considérant qu'il représente l'ensemble des propriétés empiriques des textes qui sont l'objet de la description. On peut cependant se demander pour quelle raison une norme (un sociolecte ou un idiolecte) devrait être relative à un seul niveau de description. Le choix des critères de classification paraît en partie arbitraire. En outre, ces critères ne permettent pas de sortir de classifications *ad hoc* pour établir de véritables typologies : un seul niveau permet certes le plus souvent de mettre au jour des variations et de réaliser des classifications automatiques entièrement relatives à des corpus où la variation est restreinte, mais n'ont pas permis, jusqu'à présent, de fonder des typologies. On peut faire l'hypothèse, à la fois théorique et méthodologique, qu'une norme doit être caractérisée par l'ensemble des niveaux de description et que les phénomènes qui permettraient de la définir peuvent relever de corrélations entre niveaux, plutôt que de phénomènes définis dans un seul niveau. Il est dès lors arbitraire de limiter la description à la prise en compte d'un seul niveau. En d'autres termes : pourquoi fonder des typologies textuelles sur des données qui représentent à l'évidence une part si faible des propriétés empiriques des textes ?

Cette question trouve un écho direct dans l'état de l'art en linguistique de corpus. À la limitation à des typologies « mono-niveau » sur un plan méthodologique correspond une difficulté sur le plan technique à construire les corpus nécessaires à l'investigation des relations entre niveaux. Alors que de nombreux niveaux de description peuvent aujourd'hui être annotés automatiquement (au moins les niveaux morphologiques, morphosyntaxiques, syntaxiques et lexicaux), que des progrès significatifs à court terme de ces instruments d'annotation sont peu probables, il est difficile de constituer des corpus articulant plusieurs niveaux de description. La « sortie » d'un instrument d'annotation est le plus souvent utilisée, sans autre forme d'enrichissement, comme l'objet à décrire. Cela induit une dépendance immense à l'instrument, non pas simplement vis-à-vis de ses erreurs d'annotation, ni même des choix théoriques de son concepteur mais plus profondément vis-à-vis de la nature même de l'objet empirique décrit.

Dans cet article je voudrais explorer les enjeux, pour la linguistique textuelle et les typologies textuelles, des nouveaux observables que permettent de construire les linguistiques de corpus.

1. Norme et description quantitative

La notion de norme est utile pour fonder l'utilisation de quantifications dans le cadre des linguistiques de corpus. Elle permet, d'une part, de dépasser l'opposition entre « méthodes quantitatives » et « méthodes qualitatives » et d'autre part d'articuler les moyens de description qu'apportent les linguistiques de corpus à un programme de typologie textuelle.

La norme (Coseriu, 1982) rend compte des régularités qui ne relèvent pas du système fonctionnel de la langue (d'oppositions distinctives), mais qui possèdent cependant une systématicité et caractérisent une tradition (Coseriu, 2001 : 246) :

> La norme est un un ensemble formalisé de réalisations traditionnelles ; elle comprend ce qui « existe » déjà, ce qui se trouve réalisé dans la tradition linguistique ; le système, par contre, est un ensemble de possibilité de réalisation ; il comprend aussi ce qui n'a pas été réalisé, mais qui est virtuellement existant, ce qui est « possible » [...]

L'opposition entre la norme et le système est donc une opposition entre la réalisation traditionnelle de l'activité de parler[2] et la systématicité fonctionnelle. La norme rend compte de nombreux phénomènes allant des phénomènes collocatifs[3] jusqu'aux genres et aux discours eux-mêmes[4].

2 « [...] la norme, en revanche, est un "système de réalisations obligées" [...], consacrées socialement et culturellement : elle ne correspond pas à ce qui peut se dire mais ce qui déjà "a été dit" et "se dit" traditionnellement dans la communauté considérée » (Coseriu, 2007, chap. 2) ; cf. aussi (Coseriu, 1982 :59).

3 Coseriu, 2001 : 248.

4 « [...] il est possible d'assimiler [...] les genres textuels à la *norme d'usage*, placée par Coseriu entre les pôles extrêmes de la parole et de

La norme n'est certes pas définie quantitativement, puisqu'elle rend compte d'une systématicité. La quantification n'est pas en elle-même une description si l'on ignore les unités de mesure et les objets décrits. Dans une analyse des conditions de l'utilisation des quantifications dans les sciences, Bachelard montre la nécessité d'articuler données quantitatives et objets théoriques préconstruits[5] ; de construire la mesure plutôt que de la prendre comme « intuition directe d'un objet » : « l'objectivité est [...] affirmée en deça de la mesure, en tant que méthode discursive, et non au-delà de la mesure, en tant qu'intuition directe d'un objet. Il faut réfléchir pour mesurer et non pas mesurer pour réfléchir » (1993 : 254)[6].

Cependant, si la norme n'est pas instituée par un fait de fréquence, elle est intimement liée à une dimension quantitative et des phénomènes quantitatifs sont indispensables à sa description[7] . La norme est en effet ce qui n'existe pas sans une fréquence et une attestation. Cette notion a une affinité profonde avec la dimension quantifiable des phénomènes linguistiques. Ce n'est donc pas la fréquence qui fait la norme mais au contraire la norme qui est

---

la langue » (Glessgen, 2007 : 105).

5 Le raisonnement de Bachelard concerne les sciences « dures » et ne s'applique que dans ses notions générales aux sciences humaines.

6 Canguilhem (2006 : 102) montre ainsi que, derrière les moyennes physiologiques des populations, qui peuvent passer pour un fait biologique, c'est une normativité sociale qui est à l'œuvre : « S'il est vrai que le corps humain est en un sens un produit de l'activité sociale, il n'est pas absurde de supposer que la constance de certains traits, révélé par une moyenne, dépend de la fidélité consciente ou inconsciente à certaines normes de la vie. Par suite, dans l'espèce humaine, la fréquence statistique ne traduit pas seulement une normativité vitale mais une normativité sociale. Un trait humain ne serait pas normal parce que fréquent mais fréquent parce que normal. »

7 Ce que souligne notamment Coseriu (1982 : 106, note 1, je traduis) : « L'étude statistique ou étude quantitative de la norme acquiert de plus en plus d'importance car la norme représente l'équilibre du système à un moment donné et les changements quantitatifs produisent d'habitude des changements qualitatifs : les changements dans la norme produisent des changements dans le système [fonctionnel]. »

nécessaire à l'étude de la fréquence et à l'utilisation de données quantitatives.

Bachelard souligne également les dangers du recours à des précisions quantitatives pour caractériser un objet insuffisamment défini. Des données quantitatives utilisées sans notion scientifiquement construite de l'objet décrit, présentées comme une appréhension immédiate de l'object, relève du « pittoresque »[8]. Bachelard (1993 : 255) rapporte par exemple, au XVIIIe siècle, les chiffres de Buffon, qui arriva à la conclusion qu'il y avait « 74 832 ans que la Terre avait été détachée du soleil par le choc d'une comète [...] Cette prédiction ultra précise du calcul est d'autant plus frappante que les lois physiques qui lui servent de base sont vagues et plus particulières. »

Ces exigences de tout raisonnement scientifique quant à l'usage de donnnées quantitatives doivent être poussées plus loin dans le domaine des sciences humaines – ne serait-ce que du fait du danger, précisément, d'y adopter les normes des sciences « dures »[9]. Contrairement aux sciences « dures » en effet, les points de vue ou les niveaux de descriptions ne convergent pas dans les sciences humaines, et l'objectivité ne peut consister à appréhender un même objet à partir de différents points de vue.

La question de l'utilisation non critique de données quantitatives est particulièrement cruciale pour les typologies textuelles et les classements en types de textes. En effet, les approches émergentistes des typologies textuelles sont fortement

8 « L'excès de précision, dans le règne de la quantité, correspond très exactement à l'excès du pittoresque, dans le domaine de la qualité » (1993 : 253).

9 « [...] la mentalité physiciste répandue nous a accoutumés à rechercher un autre monde « derrière » l'expérience courante et à croire que ce monde (qui justifierait le monde des phénomènes) pourrait être éventuellement découvert au moyen de l'accumulation de nombreux faits particuliers ou par les moyens instrumentaux des sciences physiques. » (Coseriu, 2007, chap. VI).

exposées aux critiques de Bachelard : on quantifie avec une grande précision des phénomènes (des classes de textes) dont on ne définit pas par ailleurs la nature, et on fait émerger par la seule mesure des objets qui restent arbitraires. Toutes les expériences concordent en effet pour montrer que, quels que soient les critères et les corpus employés, les textes se regroupent en classes[10] ; mais ces classes ne sont jamais les mêmes d'un corpus à l'autre et ces expériences ne semblent pas faire progresser vers l'identification des objets de la description. En ce sens, l'accumulation de données quantitatives est un obstacle à la cumulativité de la description.

La notion de norme est donc essentielle à un programme de typologie textuelle puisqu'elle permet d'inscrire la mesure d'une « fréquence » dans la description d'une systématicité. Elle porte précisément sur les propriétés du « réalisé », de l'attesté, en tant qu'il est distinct d'un système fonctionnel. Dans une perspective de typologie textuelle, il est utile de distinguer différents types de normes pour articuler plusieurs critères typologiques. Rastier (2001) propose pour organiser cette diversité de distinguer notamment le discours, le genre et l'idiolecte. Chaque texte est caractérisé relativement à chacun de ces types de normes. Ces différentes normes déterminent donc autant de dimensions d'un « espace des normes » dans lequel tout texte possède des coordonnés, semblables aux dimensions de l'espace variationnel.

Ces propositions permettent de souligner la nécessité de prendre en compte la pluralité et l'interaction des normes dans la description. En effet, une norme n'est jamais observable hors d'une situation d'interaction avec d'autres normes. Par exemple si l'idiolecte peut être l'objet d'une large variation au sein du discours littéraire, c'est en partie une caractérisation du discours littéraire en tant que discours : c'est le niveau du discours qui autorise la variation des idiolectes et ceux-ci, en tant qu'ils varient, caractérisent le discours littéraire et non les idiolectes eux-mêmes. De la même façon, lorsque Brunet (2004) montre que la variation entre genres surdétermine la variation entre auteurs à

10 Brunet (2004) en fait une démonstration particulièrement convaincante. Cf. également Kilgarriff (2005 : 270-271).

l'intérieur du théâtre classique (les textes d'un même genre sont classés ensemble, même s'ils appartiennent à des auteurs différents), il faut peut-être rapporter cette articulation entre idiolecte et genre au discours littéraire, voire au champ générique du théâtre classique. En tout cas, elle ne peut caractériser la relation entre genre et idiolecte dans l'absolu.

Une description d'une norme peut donc gagner à s'inscrire dans le cadre d'une architecture des normes et prendre en compte l'effet des autres normes sur les textes étudiés. En somme, il s'agit de contextualiser les normes.

Finalement, c'est l'opposition entre « méthodes quantitatives » et « méthodes qualitatives » que la notion de norme permet d'essayer de dépasser. « Quantitatif » et « qualitatif » n'opposent pas des méthodes (sinon dans un sens général, non scientifique mais plutôt seulement technique) mais des types de données. Les méthodes sont bonnes ou mauvaises en elles-mêmes indépendamment du type de données.

2. Normes et pluralité des niveaux de description

Relativement au projet de décrire quantitativement des normes, entendues comme des genres, des discours ou des idiolectes, on peut émettre l'hypothèse que des régularités sont d'autant plus définitoires et caractérisantes pour une norme qu'elles impliquent plusieurs niveaux de description et qu'elles stabilisent la relation entre niveaux de description.

En réduisant l'observation à un seul niveau de description, non seulement on réduit drastiquement les régularités observables en se passant de la combinatoire entre niveaux, mais surtout on s'interdit d'observer des régularités qui caractérisent beaucoup plus fortement une norme que celles qui s'inscrivent dans un ensemble déjà pourvu d'une systématicité fonctionnelle. Plus généralement, la distinction de niveaux de description n'a de justification que parce qu'elle permet d'observer la systématicité fonctionnelle. Il n'est donc pas justifié de se situer à l'intérieur de

l'un de ces niveaux pour caractériser la norme ; c'est au contraire par la relation qu'elle établie entre niveaux que la norme est nécessaire comme objet descriptif.

La nécessité d'accéder aux corrélations entre niveaux de description est aujourd'hui soulignée dans de nombreux secteurs de la discipline. Ainsi, depuis une perspective variationniste, Gadet (2003 : 59) écrit :

> [La] perspective qui [prend] en compte les énoncés selon des principes de différentes ordres, devrait renouveler la définition des genres en les montrant comme des faisceaux de paramètres, et non plus des rubriques rhétoriques ou situationnelles héritées de la tradition.

Dans le cadre du traitement automatique des langues, Habert & Zweigenbaum (2002 : 99) soulignent de même que « [Le traitement automatique effectif des langues] enjoint aussi de munir les données attestées d'annotations fines, multiples, permettant de progresser vers les régularités sous-jacentes. »[11].

Cette problématique est également présente dans le domaine des corpus oraux, où des dispositifs et des formats de représentation sont proposés pour permettrent l'articulation de niveaux de descriptions et de modalités (Bird & Liberman, 2001). Blache *et al.* (2007 : 1) soulignent ainsi :

> From a linguistic standpoint, language and speech analysis are based on studies of distinct research fields, such as phonetics, phonemics, syntax, semantics, pragmatics or gesture studies. [...]. The perspective adopted by modern linguistics is a considerably

11 A partir du niveau morphologique, Baayen (1994 :32) souligne également la nécessité de croiser les niveaux : « this suggests that the combined quantitative analysis of morphology on the one hande [...] and syntax and pragmatics on the other [...] constitutes a robust and fruitful line of inquiry into the sociolinguistic and stylistic aspects of language use. » En apprentissage, Yvon (2006 : 34) note : « [...] la description linguistique fournit des descripteurs de plusieurs niveaux, qu'il peux être souhaitable d'utiliser conjointement. »

> broader one: even though each domain reveals a certain degree of autonomy, it cannot be accounted for independently from its interactions with the other domains. Accordingly, the study of the interaction between the fields appears to be as important as the study of each distinct field.

Ou Blanche-Benveniste (2005 : 47) :

> On peut enfin produire des analyses distributionnelles étendues à de vastes contextes, sur lesquels on peut mesurer l'effet des collocations entre prosodie, lexique et grammaire.

Enfin, dans le domaine diachronique, Coseriu (2007, chap. IV) souligne cette interdépendance entre niveaux : « [...] il y a une intime solidarité entre le phonétique, le lexical et le grammatical ; ce qui, dans la perspective diachronique, signifie qu'un changement affectant n'importe lequel de ces aspects possède des répercussions sur l'ensemble du système. »[12]

Pour la description de normes en corpus, l'accès aux corrélations entre niveaux de description permet de proposer de nouveaux observables voire de nouvelles catégories descriptives. En effet, les catégories descriptives de la linguistique textuelle, comme les notions d'isotopies, de rythme, les catégories de l'analyse actantielle, ou la modélisation du contexte, permettent de caractériser le fonctionnement sémantique profond d'un texte. Or, ces catégories descriptives peuvent aujourd'hui être constituées en observables, en corpus, et être utilisées dans une perspective typologique.

Dans (Loiseau, 2007) j'ai proposé plusieurs observables inspirés des catégories descriptives de la linguistique textuelle pour caractériser par des phénomènes quantitatifs le fonctionnement d'un discours.

Par exemple, dans l'œuvre de Gilles Deleuze, la prise en compte de la linéarité (la « tactique ») de phénomènes (lexicaux comme grammaticaux) a permis de caractériser le paragraphe

12 Cf. également Glessgen, 2007 : 121.

comme unité : le début et la fin des paragraphes opposent de manière récurrente et significative des registres, des acceptions de certains lexèmes, et des orientations axiologiques. Le paragraphe, peu pris en charge et au statut controversé, a pu être caractérisé dans le corpus étudié comme un palier sémantiquement fort : l'enjeu est important, puisqu'il s'agit là de qualifier de nouvelles unités linguistiques, notamment supra-phrastique, dans le contexte d'une norme.

Un autre observable a essayé de caractériser des textes par leur structure isotopique (par un aspect du fonctionnement de leur contexte). Pour cela j'ai représenté par un graphe les relations de cooccurrences dans un contexte fortement assimilateur (dépendances à la conjonction *ou*). La dépendance à un *ou* est utilisée pour sa « valeur contextuelle » : quelles que soient les acceptions de *ou* dans ses différentes occurrences, le contexte de la dépendance à *ou* est à chaque fois fortement assimilateur et implique l'existence d'afférences sémantiques fortes entre les lexèmes coordonnés. Dans le graphe construit, les noeuds représentent l'ensemble des lexèmes dans la dépendance d'un *ou* et les arêtes relient deux lexèmes qui cooccurrent au moins une fois dans la dépendances à une même occurrence de la conjonction. Par une classification automatique de ce graphe, on obtient des ensembles lexicaux fortement interdéfinis (constitués de noms qui cooccurrent dans les contextes de la conjonction). Ces ensembles partagent des traits sémantiques très abstraits, dits dimensionnels (tels que /pluralité/, ou /borné/). Cette abstraction des sèmes isotopants caractérisent sans doute le discours philosophique : l'abstraction conceptuelle trouve une correspondance dans des mécanismes d'abstraction sémantique par neutralisation des sèmes domaniaux[13]. Ici, la méthodologie permet de caractériser des textes non pas directement par leur unité mais par la nature des relations contextuelles établies entre certaines unités. Le fonctionnement des contextes locaux est caractérisé à une échelle « globale ».

13 C'est-à-dire des sèmes qui indexent dans un domaine de type lexicographique.

Enfin, l'ensemble de l'axe diachronique de l'œuvre de Gilles Deleuze a été appréhendé par une analyse factorielle multidimentionnelle regroupant plusieurs niveaux de description (morphologique, syntaxique, morphosyntaxique, lexicaux, typographique). L'ensemble des niveaux de description contribue à opposer une première période académique[14], une période centrale de l'engagement politique[15], et une dernière période caractérisée par un replis littéraire[16].

Ainsi, des observables articulant méthodes quantitatives et linguistiques textuelles peuvent contribuer à caractérier les textes au-delà de profils morphosyntaxiques et lexicaux. Cela induit une notion de contextualité entre niveaux de description.

3. Typologies textuelles et typologies linguistiques

On peut essayer de tirer toutes les conséquences, dans la perspective ouverte par la notion de norme, de l'homologation entre le système fonctionnel et le système « normal » : dans cette perspective, on peut faire l'hypothèse qu'il n'y a pas de solution de continuité entre la langue et les sociolectes ou l'idiolecte et qu'il n'y a pas de différence de nature entre la diversité des langues et la diversité des normes.

Si l'hypothèse de l'absence de solution de continuité entre diversité des langues et diversité des genres et des discours est peut-être prématurée en l'état actuel des typologies textuelles, il y a sans doute néanmoins matière à s'interroger sur les intersections au moins méthodologiques entre les différents cadres typologiques. Les questions qui apparaissent aujourd'hui en linguistique de corpus font écho à des alternatives déjà posées dans le cadre des typologies linguistiques – « caractérologie » ou identification des langues, typologie globale ou relative à des

14 Caractérisée par le *nous*, les phrases complexes, l'apparat critique (notes, divisions), les temps du subjonctif, etc.

15 Caractérisée par l'impératif, le point d'exclamation, la 2nde personne, etc.

16 Caractérisée par la première personne du singulier, l'indéfini, les noms propres, le pluriel, les points de suspension, l'imparfait, etc.

critères isolés[17] – et cette parenté ne doit sans doute pas être ignorée. Coseriu (2001 : 242) souligne lui-même cette absence de solution de continuité entre langue et norme :

> À la rigueur il n'y a pas, à cet égard, de différence essentielle entre deux techniques du discours à l'intérieur d'une langue historique et deux langues historiques différentes. La différence est tout simplement de degré de diversité : à l'intérieur d'une langue historique, les différences sont moindres qu'entre cette même langue et une autre langue historique et, normalement, elles n'affectent pas tout le système phonologique toute la grammaire et tout le système lexical, mais, selon les cas, des sections plus ou moins étendues de ces systèmes (cependant, entre deux langues historiques différentes, les différences peuvent être moindres que, par exemple, les différences entre deux « dialectes » d'une troisième langue historique).

Depuis une perspective de typologie linguistique, Hagège (2005 : 68) en vient à s'interroger sur la nécessité, pour des typologies de langues, de prendre en compte le discours : « la typologie ne saurait se contenter d'étudier la grammaire. Le discours est une réalité fondamentale, et c'est au croisement des deux que devraient se situer les études typologiques. » Or, prendre en compte le discours implique de s'appuyer sur une typologie des textes.

Plus généralement, la notion de langue ne pose pas moins de difficulté que celle de sociolecte[18], et leur distinction relève peut être encore, avant tout, du « sens commun »[19]. En ce sens, c'est le postulat d'une identité de nature entre ce qu'on désigne comme diversité des langues et diversité des normes qui est le postulat le

17 Cf. Shibatani & Bynon (1997).

18 « Dans le continuum de la communication, nulle césure ne permet d'extraire des entités stables auxquelles on pourrait assigner le nom de langue. » (Laks, 2002 : 33).

19 « Pour le sens commun, on ne saurait en effet douter un seul instant que les langues existent, constituent des objets définissables et circonscriptibles, dans le temps comme dans l'espace géographique et social » (Laks, 2002 : 28)

moins coûteux et c'est seulement à la condition de ne pas les opposer prématurément que l'on peut espérer les distinguer rigoureusement. Enfin, peut-être la définition de modèles typologiques plus adéquats pourrait-elle bénéficier d'un rapprochement des typologies linguistiques et des typologies textuelles[20].

La prise en compte de la diversité des normes est également présente, dans le domaine de la description des langues, à travers l'apparition du paradigme de la « documentation des langues » (Himmelmann, 2002)[21]. Il s'agit d'insister sur la nécessité, à côté de la description proprement dite, de documenter et archiver les langues décrites dans de vastes corpus raisonnés. Ces corpus doivent permettre, d'une part, de nourrir et vérifier les hypothèses linguistiques elles-mêmes, et d'autre part de sauvegarder voire « revitaliser » les langues en danger. Dans ce cadre, la diversité des usages reçoit une attention croissante (Himmelmann, 2002 : 19), puisque la documentation d'une langue, au-delà de la description de son système fonctionnel, ne saurait ignorer ses usages :

> [...] many of the insights gained in the literature on genre are relevant to the problem of determining the kind and number of communicative events to be included in a language documentation.

Cette convergence d'intérêt désigne sans aucun doute l'élaboration d'un véritable cadre théorique et méthodologique pour les typologies textuelles comme un enjeu majeur pour la

20 « [...] il n'existe pas de structuration hiérarchique cohérente possible pour tous les genres textuels. » (Glessgen, 2007 : 107) ; « Les français, et avec eux Saussure et une grande partie des jeunes linguistes du temps, contestent tout ensemble l'existence de langue et de dialectes nettement séparés dans le temps et dans l'espace, l'existence de familles linguistiques distinctes et partant, récusent les filiations les plus reçues de la généalogie linguistique. » (Laks, 2002 : 27).

21 Plusieurs intersections avec les linguistiques de corpus émergent dans ce cadre (Bird & Simons, 2003).

linguistique. Les nouveaux moyens de description de la textualité doivent donc sans doute être articulés à un programme de typologie textuelle, au-delà des perspectives de classification automatique. Il semble que le paradigme typologique est le cadre scientifique dont relève un projet de description de la diversité des normes linguistiques. C'est notamment à travers leur contribution à un programme typologique que peuvent être évalués les nouveaux observables produits par les linguistiques de corpus.

Conclusion

Les considérations méthodologiques de cet article peuvent être résumées par la proposition de considérer deux nouveaux types de contextualité : la contextualité comme interaction des différents types de normes et la contextualité comme interaction des niveaux de description. Elles consistent également à essayer d'articuler linguistique textuelle et méthodes quantitatives pour proposer de nouveaux moyens de décrire la textualité, notamment en dépassant l'utilisation d'un seul niveau de description, lexical ou morphosyntaxique, pour accéder, en corpus, à des solidarités entre niveaux de description, et à d'autres types de contextualité. Enfin, elles consistent à articuler ces nouveaux moyens de description de la textualité à un programme de typologie textuelle, au-delà des perspectives de classification automatique. Celui-ci, pour construire ses objets, peut s'attacher à rendre compte de l'articulation des normes dans un texte. Les observables ainsi construits renouvellent profondément les capacités de caractérisation quantitative des textes et donc le matériau d'un travail comparatiste.

Textes cités

Bachelard G. (1993 [1938]) *La Formation de l'esprit scientifique*, Paris : Vrin.

Baayen H. (1994) "Derivational productivity and text typology", *Journal of Quantitative Linguistics*, 1, pp. 16-34.

Biber D. (1988) *Variation across speech and writing*, Cambridge : Cambridge University Press.

Bird S. & Liberman M. (2001) "A formal framework for linguistic annotation", *Speech Communication*, 33-1/2, pp. 23-60.

Bird, S. & Simons G. (2003) "Seven Dimensions of Portability for Language Documentation and Description", *Language*, 79, pp. 557-582.

Blache P., Rauzy S., Ferré G. (2007) « An XML Coding Scheme for Multimodal Corpus Annotation », *Proceedings of* Corpus Linguistics.

Blanche-Benveniste C. (2005) « L'Étude grammaticale des corpus de langue parlée en français » *in* Williams G (éd.) *Les Linguistiques de corpus*, Rennes : Presses Universitaires de Rennes, pp. 47-66.

Brunet, E. (2004) « Où l'on mesure la distance entre les distances », *Texto !*, mars 2004. En ligne : <http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html>.

Canguilhem G. (2006 [1966]) *Le Normal et le pathologique*, Paris : PUF.

Coseriu E. (1982 [1952]) « Sistema, norma y habla » in *Teoria del lenguaje y lingüistica general*, Madrid : Gredos.

Coseriu E. (textes réunis par Dupuy-Engelhardt H., Durafour J.-P., Rastier F.) (2001) *L'Homme et son langage*, Leuven : Peeters.

Coseriu E. (trad. T. Verjans) (2007 [1958]) *Synchronie, diachronie et histoire*, Paris : Texto ! (non paginé).

Gadet F. (2003) *La Variation sociale en français*, Paris : Ophrys.

Glessgen M.-D. (2007) *Linguistique romane – Domaines et méthodes en linguistique française et romane*, Paris : Armand Colin.

Habert B. & Zweigenbaum P. (2002) « Régler les règles » *TAL*, 43-3, pp. 83-105.

Hagège C. (2005) « De la place réelle de la transitivité, ou la typologie linguistique entre passé et avenir » *in* Lazard G., Moyse-Faurie C. (éd.) *Linguistique typologique*, Paris : Presses Universitaires du Septentrion, pp. 55-69.

Himmelmann N. « Documentary and descriptive linguistics », *Linguistics*, 36, pp. 161-195.

Kilgarriff A. (2005) « Language is never, ever, ever, random », *Corpus Linguistics and Lingustic Theory*, vol. 1, pp. 263-276.

Laks B. (2002) « Le Comparatisme :de la généalogie à la génétique, *Langage*, 146, pp. 19-45.

Loiseau S. (2006) *Sémantique du discours philosophique chez Deleuze : du corpus aux normes*, Thèse de doctorat, Université Paris X-Nanterre.

Loiseau S. (2007) « CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations », *Corpus*, n°6, Pincemin B. (éd.) « Contexte, interprétation, codage », pp. 153-186.

Loiseau S. (à paraître) « Construction et interrogation de corpus multi-annotés » *TAL*, 49-2.

Loureda Lamas Ó. (2004) « Pasado, presente y futuro de la tipologia textual », in Hassler G. & Volkmann G. (éd.) *History of Linguistics in Texts and Concepts*, Münster : Nodus Publikationen.

Rastier F. (2001) *Arts et sciences du texte*, Paris : PUF.

Shibatani M. & Bynon T. (1997) *Approaches to Language Typology*, Oxford: Oxford University Press.

Yvon F. (2006) *Des apprentis pour le traitement automatique des langues*, Habilitation à diriger des recherches, Paris : Université Pierre et Marie Curie-Paris VI.